基于 WMF_LDA 主题模型的文本相似度计算 *

张 璐[a]，芦天亮[a, b]，杜彦辉[a, b]

(中国人民公安大学 a. 信息技术与网络安全学院; b. 网络空间安全与法治协同创新中心, 北京 100038)

摘 要：文本相似度的判断和计算是自然语言处理领域中具有重要意义和研究价值的一部分内容。利用 LDA 模型进行文本相似度的计算考虑到了语义特征，但是存在词语数量多、未结合词语语义、未从文本层面挖掘和利用不同类别文本固有的领域间差异的缺点。针对以上问题，提出 WMF_LDA(词语合并与过滤潜在狄利克雷分布)主题模型。将领域词和近义词进行统一化映射，并根据词性将文本进行过滤，最后再进行主题建模。实验证明，该方法使得建模时词语量大大减少，减少了建模过程的时间消耗，提高了最后的文本聚类的速度。并且与其他文本相似度方法相比，本文提出的方法在准确度上也有一定程度的提升。

关键词：词语语义；词语合并；词性筛选；文本相似度

中图分类号：TP391.1   doi: 10.3969/j.issn.1001-3695.2018.04.0219

Text similarity calculation based on WMF_LDA topic model

Zhang Lu[a], Lu Tianliang[a, b], Du Yanhui[a, b]

(*a. Information Technology & Network Security Institute*, *b. CIC of Security & Law for Cyberspace*, *People's Public Security University of China*, *Beijing* 100038, *China*)

Abstract: Text similarity calculation is a significant part with great research value in the field of NLP (Natural Language Processing) . The calculation of text similarity with LDA (Latent Dirichlet Allocation) model takes into account the semantic features, but it has the disadvantages of a large number of words, unconformity of the semantics of words, and the inability to dig and exploit the inter-domain differences inherent in texts of different categories. This paper proposes WMF_LDA topic model (Word Merging and Filtering_LDA) . This model maps domain words and synonyms, and filters the words based on POS. Finally, LDA theme is used on the processed result. Experiments show that this method greatly reduces the amount of words during modeling, reduces the time consumption of the modeling process, and improves the speed of the final text clustering. And compared with other text similarity methods, the method proposed in this paper also has a certain degree of improvement in accuracy.

Key words: word semantics; word merging; POS(Part-of-Speech) filtering; text similarity

0 引言

文本相似度是在语言学、心理学和信息理论等领域内均被广泛研究的重要课题[错误!未找到引用源。]。尤其在自然语言处理方面，文本相似度计算是其中的重要研究内容和研究方向。在信息检索和比对方面，文本相似度计算为其提供手段和方法，良好的相似度计算算法可以很好的提高，甚至很大程度上决定了信息检索和比对结果的精确度[1]。文本相似度的应用范围非常广泛，在图像检索领域，利用图像周围文字的相似度程度，可以进一步判定其图像的相似度情况，可以获得更好地检索精度；在文本聚类方面，文本相似度算法为其提供了依据，从根本上决定了文本聚类的结果和精确度。除此之外，文本相似度计算还应用于文本摘要生成[错误!未找到引用源。]、文档重复度检测等领域[错误!未找到引用源。]。

1 相关工作

一直以来，文本相似度研究都是自然语言处理的重要研究课题。传统的 VSM 方法以 TF-IDF 作为特征构建向量，并以余弦距离计算文档的相似度[错误!未找到引用源。]，但是这种方法单纯以词频作为特征，没有考虑词语和文本的语义特征。苏小虎等人[错误!未找到引用源。]结合原有特征项权重和文档中特征项自身的领域权重，改进传统 VSM 方法。黄承慧等人[错误!未找到引用源。]提出词项相似度加

收稿日期：2018-04-20；**修回日期：**2018-05-16 **基金项目：**国家重点研发计划重点专项项目（2017YFB0802804）；国家自然科学基金资助项目（61602489）；中国人民公安大学 2018 年基本科研业务费科研机构项目（2018JKF504）

作者简介：张璐（1994-），男，硕士研究生，主要研究方向为自然语言处理（gadxyjsyzl@163.com）；芦天亮（1985-），男，副教授，博士，主要研究方向为信息安全；杜彦辉（1969-），男，教授，博士，主要研究方向为信息安全．

权树，将词语相似度映射到文本相似度，但是存在计算量大的缺点。谷重阳错误!未找到引用源。利用计算出各词项的 TF-IDF 值对相似度计算公式进行了改进，将词汇的相似度作为权值对余弦距离公式进行改进。Blanco 等人**错误!未找到引用源。**提出一种新的句型和语法的分析方法，从句子中抽取语义关系，并进行文本相似度的计算。Atoum 等人错误!未找到引用源。利用距离和内容计算词语相似度，并通过加权的方式扩展到文本相似度。在短文本相似度计算方面，黄贤英等人提出按照词性对文本中出现的所有词项进行分类，并按照重要程度对不同词性赋予不同的权值错误!未找到引用源。错误!未找到引用源。。

在神经网络和深度学习方面，黄江平等人错误!未找到引用源。提出基于卷积神经网络 CNN 的文本相似度检测模型。Kenter 等人错误!未找到引用源。综合了不同条件下获得的不同维度的词向量，并将词语相似度映射到文本相似度。Kusner 等人错误!未找到引用源。通过词移距离（word mover's distance, WMD），利用词向量计算文本相似度。Neculoiu 等人错误!未找到引用源。利用 LSTM 框架，获取不等长字符串之间的语义相似性。Kashyap 等人结合了文本的潜在语义和机器学习，综合了多种语言资源的数据错误!未找到引用源。。

以文本主题为切入点，也是计算文本相似度的一种方法。孙昌年等人利用 LDA 对文本进行建模，利用主题差异表示文本的相似性，但是这种方法存在词语规模大，建模速度慢的缺点错误!未找到引用源。。张超等人结合词性改进了 LDA 算法，一定程度上减少了词语规模，提高了建模速度，但是并没能进一步结合词语之间的语义关系，挖掘不同领域内的文本之间固有的差异性错误!未找到引用源。。本文针对以上利用 LDA 进行文本相似度计算时的缺点，提出了 WMF_LDA 主题模型，结合词语语义和词性信息，利用文本集之间的领域差异，改进了传统的 LDA 模型在文本相似度计算领域的应用。

2　WMF_LDA 主题模型工作过程

2.1　模型结构

不同类型的文本具有其本身固有的与其他类别文本的差异性，这种差异性主要体现在用词方面。不同类型的文本，都有一套在其领域内常用的词语列表，我们称之为领域词表。领域词表中的词语被称为该类型文本所对应的领域词。所提出的 WMF_LDA 模型就是在原有 LDA 的模型基础上，最大限度的利用这种不同类型文本在领域词方面上的差异性。WMF_LDA 主题模型的工作流程如图 1 所示。

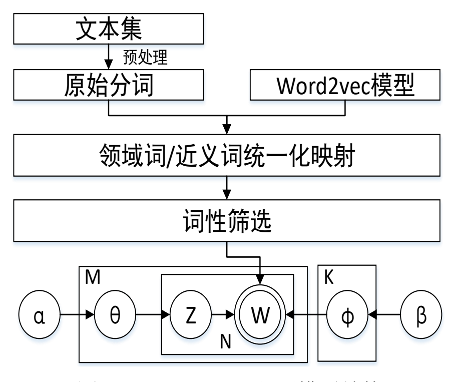


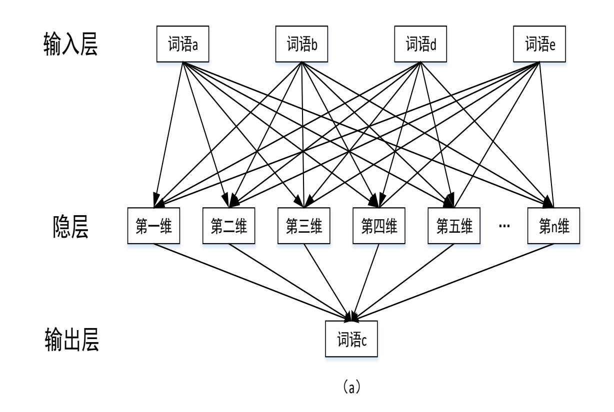
图 1　WMF_LDA 主题模型结构

在 WMF_LDA 主题模型中，对于原文文本集，按照正常 LDA 模型工作之前的预处理过程将其进行分词处理；然后，根据事先训练好 word2vec 词向量模型，在语义层面上将领域词和近义词进行统一化的映射；然后，根据文本中名词和动词对于文章语义结构影响较大的特点，将映射后的词语集按照词性进行筛选，保留名词和动词，将其他词性的词语过滤掉；最后，对经过以上处理之后的结果进行 LDA 主题建模。

在图 1 中，K 为预设的文档的主题数，M 为语料库中包含所有的文档数量，N 表示语料库中所含的全部词语，W 表示可被观测到的词项，Z 表示所选定的该词语的所属主题，θ为文档-主题的概率分布，φ为主题-词语概率分布，α为θ分布的超参数，β为φ的超参数。

2.2　词语相似度计算

本文中采用 word2vec 模型进行词语的向量化表示。其基本思想是根据词语在文章中的位置，综合了上下词信息来计算获取词向量，因此计算出来的词向量将具备一定程度的语义信息。其包括两种训练模型，CBOW（continuous bag-of-words model）和 skip-gram（continuous skip-gram model）。



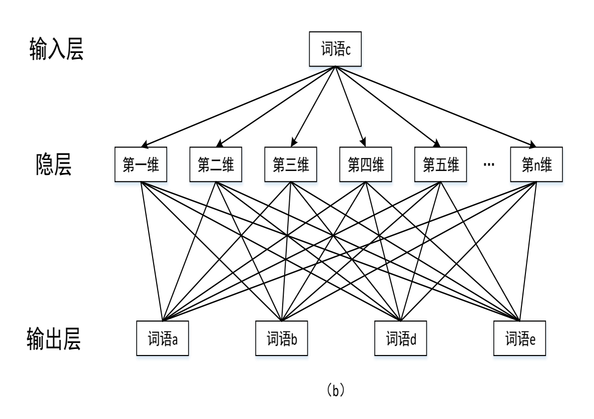


图 2　word2vec 训练模型

图 2（a）表示 word2vec 的 CBOW 模型，其通过某个词的上下词语来获取该词的向量表示；图 2（b）表示 word2vec 的 Skip-gram 模型，通过将某词映射为其上下文临近词来获取该词语的向量表示。

本文通过以上 word2vec 模型，将每个词语表示为 N 维的词向量，词语之间的相似度则通过如下的余弦夹角来进行计算和表示。

$$\mathrm{Sim}(w_1,w_2)=\frac{\sum_{i=0}^{N}w_{1i}w_{2i}}{\sqrt{(\sum_{i=0}^{N}w_{1i}^2)(\sum_{i=0}^{N}w_{2i}^2)}}\quad(1)$$

2.3　基于语义的词语合并

不同类别的新闻文本具有其常用的词语集合，或专业领域的词语集。本文采用复旦语料库从中选择太空、艺术、农业、经济、政治五个类别中的随机 200 篇文章，计算不同词语在不同类别文本中的出现频率，结果如表 1 所示。“灌溉”“农村”等词语在“农业”类别中出现的次数较多，而在其他类别的文本中出现次数较少甚至不会出现；同理，“钢琴”“航天”等词语则分别在“艺术”与“太空”两类文本集中出现的次数较多。

表 1　不同词语在不同类别文本中的出现次数

	太空	艺术	农业	经济	政治
灌溉	0	2	65	5	0
农村	0	59	2306	476	165
钢琴	0	23	0	0	0
作品	0	1800	0	1	9
航空	226	0	1	6	4
航天	100	1	1	10	0
经济	25	240	2169	6575	1774
货币	0	6	50	433	18
省委	0	0	5	1	58
政治	0	648	96	432	11189

因此，我们可以得知，不同类别的新闻文本都有其各自的领域词集，其中的词语在该类别的新闻中出现次数较多，而在其它类别的文本中出现次数较少。基于以上分析，我们可以提出如下假设：

假设 1　不同类别文本有其相对固定、区别于其他类别文本的、体现其领域专业性的词语集合。

假设 2　若两词属同一领域词集，则其相似程度较其属不同领域词集更大。

除此之外，文本在之后步骤中 LDA 模型进行主题建模，首先是将原始文本转换为词频矩阵，因此我们将同义词和领域词映射到同一个词语表达上可以增加该领域的独特性，提高领域词的出现频率。另外，我们采用 Gibbs 采样，得到所有词的主题分布，统计某篇文档中所有词语的主题计数，便可得到该文档对应的主题分布；同理，统计所有语料库中所有词的主题计数，便可得到各个主题对应的词语分布。因此可以得到如下推断：

推断 1　某篇文章中包括某个主题的词语的数量越多，该文章包含该主题的概率则越大。

基于以上假设与推断，提出本文思路：通过相似度的计算将同（近）义词与同一专业领域内的词语映射到一个词语上，如将“航空”“航天”“宇航”等词语统一映射为“航空”。这样可以最大限度的发挥各类文本领域词集内的词语对该类别文本的标志作用，同时通过将同义词映射为统一词语表示，提高了该词语的出现频率，在通过 Gibbs 采样获取文本—主题概率分布时，提高文本在该主题下的分布概率值。

另外，通过设定词语之间的相似度阈值 t，来判定词语是否相似，从而进行统一化的映射，可以大幅的降低词语规模，提高了 LDA 建模效率。

2.4　基于词性的词语筛选

根据中文文本的特点，在文本的语义结构上，名词和动词对于一篇文章的内容理解、语义结构等方面均具有重要的作用，将文本中除名词和动词之外的词语删除并不会影响我们对于整篇文本语义的把握和理解。此外，在文本的组成结构方面，名词和动词也占总体词语数量的比重也较大，对于文本的结构组成其重要作用。

无论在语义结构还是组成结构，名词和动词都是一篇文本的核心要素。因此，文本针对此特点，将通过上一步进行词语合并与映射之后的结果，按照词性进行筛选，保留对文本语义和结构影响较大的名词和动词，而过滤掉其他影响较小的词语，排除助词、语气词等无关词语对后续建模工作的影响，进一步降低词语规模。

2.5　WMF_LDA 主题建模与采样

本文提出的 WMF_LDA 模型在建模阶段采用原始的 LDA 模型，其基于这样的假设：每篇文章包含若干主题，其出现概率不同，同时不同主题下包含若干词语，同一主题下不同词语的出现概率也不尽相同，即一篇文章是由多个主题以某种分布式概率构成，而各主题则是由各项词语以某种分布式概率构成，而忽略掉词语的语法结构和出现的先后顺序[错误!未找到引用源。]。因此，

对于 LDA 来说，文章由主题构成，主题由词构成，而文章—主题的分布$\vec{\theta}$与主题—词语的分布$\vec{\varphi}$均服从多项分布。结合图 1 中的 LDA 建模过程，可以用联合公式表示第 m 篇文章的生成过程：

$$\mathrm{P}(\overrightarrow{Z_m},\overrightarrow{W_m},\overrightarrow{\theta_m},\varphi|\vec{\alpha},\vec{\beta}) = \prod_n^N P(W_{m,n}|\overrightarrow{\varphi_{Z_{m,n}}})P(Z_{m,n}|\overrightarrow{\theta_m})P(\overrightarrow{\theta_m}|\vec{\alpha})P(\varphi|\vec{\beta}) \quad (2)$$

通过以上公式可知，对于第 m 篇文章的生成，本质上是通过循环生成每一个词的过程。对于第 m 篇文章的第 n 个词语的生成，则其具体算法步骤如下：

a)以α作为超参数，通过狄利克雷分布获取文档-主题概率分布$\overrightarrow{\theta_m}$。

b)对获取的文档-主题概率分布$\overrightarrow{\theta_m}$，通过多项式分布，获取该词所属的主题$Z_{m,n}$。

c)以β为超参数，结合该词所属的主题$Z_{m,n}$，通过狄利克雷分布获取主题-词语概率分布$\overrightarrow{\varphi_{Z_{m,n}}}$。

d)对获取的主题-词语概率分布$\overrightarrow{\varphi_{Z_{m,n}}}$，通过多项式分布，获取该词$W_{m,n}$。

e)重复以上步骤 a)~d)N_m次，生成第 m 篇文章的N_m次。

f)重复以上 a)~e)步骤 M 次，生成 M 篇文章。

模型中最主要的需要求解的两个参数为$\overrightarrow{\theta_m}$与$\overrightarrow{\varphi_{Z_{m,n}}}$，即文本-主题与主题-词语这两个多项分布。在 WMF_LDA 主题建模过程中，α和β为需要提前进行确定的超参数。在文本中，α和β取经验值：$\alpha = 50/\mathrm{K}$，$\beta = 0.01$。

除此之外，上式中的$Z_{m,n}$也是未知的，因此需要根据已经生成的文本中词语分布，倒推得到需要的参数分布，即 Gibbs 采样的方式获取需要的参数。文本 WMF_LDA 模型所采用的采样流程如下：

a）获取经过以上词语合并与筛选处理后的词语集合，对其中的每一个词语随机初始化一个主题$z^{(0)}$。

b）对每一个词语，根据如下 Gibbs 采样公式更新当前词语的主题概率，即排除当前词语的主题分配，根据其他所有词的主题分配重新估计当前词语在各个主题下的概率。其中，$n_t^{(k)}$表示主题 k 下出现词语 t 的数量，$n_m^{(k)}$表示文档 m 中出现主题 k 的次数，¬i表示除去下标为 i 的词。

$$\mathrm{P}(z_i = \mathrm{k}|\overrightarrow{z_{\neg i}},\vec{\mathrm{w}}) \propto \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t}(n_{m,\neg i}^{(k)} + \alpha_k) \quad (3)$$

c）重复以上过程直至采样收敛。

d）通过以下公式计算得文本—主题概率分布情况

$$\theta_{m,k} = \frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K} n_m^{(k)}+\alpha_k} \quad (4)$$

2.6 文本相似度计算

通过以上 WMF_LDA 主题模型，对于语料库中的每一篇文本，得到了其在主题层面上的概率分布。本文以文本之间主题概率差异表示文本的相似程度，因此选择相对熵（KL 距离）作为文本相似程度的判定标准。同时，由于 KL 距离是非对称的，本文采用其变种， JS 距离进行文本相似程度计算，其计算公式如下，其中d_1、d_2表示经过以上 WMF_LDA 建模得到的文本概率分布。

$$\mathrm{Sim}(d_1,d_2) = \frac{1}{2}\left[D_{kl}\left(d_1,\frac{d_1+d_2}{2}\right) + D_{kl}\left(d_2,\frac{d_1+d_2}{2}\right)\right] \quad (5)$$

$$D_{kl}(d_1,d_2) = \sum_{i=1}^{N}(d_{1i}\log\frac{d_{1i}}{d_{2i}}) \quad (6)$$

3 实验结果与分析

3.1 实验数据

本文所采用的实验数据分为 word2vec 词向量训练和利用 WMF_LDA 主题模型进行建模与文本相似度计算两部分。

进行词向量训练时，综合采用复旦大学语料库、腾讯新闻语料库、搜狗实验室新闻语料库、凤凰新闻网、网易新闻语料库、人民网、维基百科等多个中文文本语料库，共 2813611 篇新闻文本，83 万词条。

进行 LDA 的建模与文本相似度计算时，采用的是复旦大学语料库的部分文本数据。本文选取其中艺术、太空、农业、经济和政治五类语料，每类随机选择 200 篇文本，共计 1000 篇文本进行建模与相似度计算。

3.2 文本聚类与相似度衡量

利用本文提出的文本相似度计算方法，计算出两篇文章在主题分布上的相似程度，并以此作为文本之间距离，对测试集中全部样本进行文本聚类。根据聚类结果，判断每一篇文章是否被划分至正确的类别，同时判断每一个类别是否包含对应属于该类别的文本。综合判定本文提出的文本相似度计算方法的准确性。

上述聚类结果的准确程度通过 F1 值进行衡量。F1 值是在机器学习、自然语言处理、信息检索等领域进行评估的重要指标。根据聚类结果，计算聚类 j 所属类别 i 的查准率P(i,j)，与聚类 j 所属类别 i 的找回率R(i,j)[1]。其计算公式如下所示：

$$\mathrm{P(i,j)} = \frac{n_{ij}}{n_j}, R(i,j) = \frac{n_{ij}}{n_i} \quad (7)$$

其中：n_{ij}表示聚类结果为 j 的文本中属于类别 i 的文本数量；n_i表示类别为 i 的文本数量；n_j为聚类结果为 j 的文本数量。

通过P(i,j)与R(i,j)根据以下公式计算得到其对应 F 值：

$$\mathrm{F(i,j)} = \frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)} \quad (8)$$

全局聚类的 F1 值的计算公式如下：

$$\mathrm{F1} = \sum_i^N \frac{n_i}{n} max_j(F(i,j)) \quad (9)$$

其中：N 为测试集中包含的文本类别数量，n 表示测试集中文本数量。全局聚类的 F1 值越大，说明聚类效果越好，反映出文本相似度计算算法效果越好。

3.3　基于语义的词语合并

在本文进行实验所采用的 1000 篇文本中，共包涵 6 万多个不同的词语。利用 word2vec 模型在语义层面上对词语进行合并和统一化映射时，本文设定阈值$t = 0.5$，对大于该阈值的词组进行合并。最终将词语数量降低为 4 万，仅为原来的三分之二，可以有效地提高后续进行 LDA 建模的速度。表 2 展示了文本集中的一些词语在根据语义信息与其他词语进行合并之后结果。

表 2　文本集中部分词语映射后结果

统一映射后的词语	原始文本集中的词语
国际航空	国际航空、航空、航空公司
增容费	增容费、入网费、配套费、电费、安装费、煤气费
高宗	高宗、肃宗、明宗、仁宗、孝宗、太宗、成宗
横斜	横斜、疏影、青绿、暗香、草绿
悬臂梁	悬臂梁、悬臂、铣刀、摇臂、腹板、工字钢
中央组织部	中央组织部、团中央、中央宣传部、总政治部
倒茬	倒茬、轮作、田块、密植、翻耕、黄萎病
商品量	商品量、收购量、商品率、生产量、储备量
广告宣传	广告宣传、广告、商业广告、宣传、虚假
通俗歌曲	通俗歌曲、英文歌曲、流行歌、爵士乐、革命歌曲

如表 2 所示，“国际航空”“航空”“航空公司”明显属于“太空（Space）”类别的词语，而在其他类别中很少出现，因此统一映射为“国际航空”。同理，“中央组织部”“团中央”“中央宣传部”等词语则在“政治（politics）”类中出现的次数较多，则统一映射为“中央组织部”。通过表 2 可得，根据词语语义进行领域内词语的合并，可以提高该领域词的出现频次，提高其对所属领域的反映和代表能力。

3.4　LDA 主题数量选择

在对文本集进行 LDA 主题建模之前，需要事先确定主题数量 K。K 值选择较小，则无法将不同主题进行区分，会出现多个主题映射到同一个主题上的情况，无法准确的通过主题分布的差异表示计算出文本相似程度；K 值选择较大，则意味将每一篇文本映射到多个不同的主题维度上，忽略了即使相同类型的文本在主题的细节上也会存在的差异性。同时，过多的维度也会增加后续的计算时间，降低计算效率。因此，不同 K 值的选择将直接影响到后边 LDA 模型的准确度。

因此，在进行最终准确率实验进行比较之前，需要首先确定建模过程中所采用的 K 值。具体过程算法如下：

a)针对 3.1 节描述的 1000 篇测试集文章，利用上述 2.3 和 2.4 描述的词语合并与筛选过程对其进行处理。

b)确定建模参数。其中，α和β采取经验值，并确定待测定的 K 值范围。对于每一个 K 值，计算下列过程：

c)根据上述 2.5 节进行主题建模，获取 1000 篇测试集文本在 K 个维度上的主题分布。

d)根据上述 2.6 节内容，计算 1000 篇测试即文本中两两之间的相似度值。

e)根据上述 3.2 节内容，利用 K-means 聚类算法，对 1000 篇测试集文本进行聚类计算。

f)根据聚类结果，计算该 K 值对应下的全局准确度 F1。

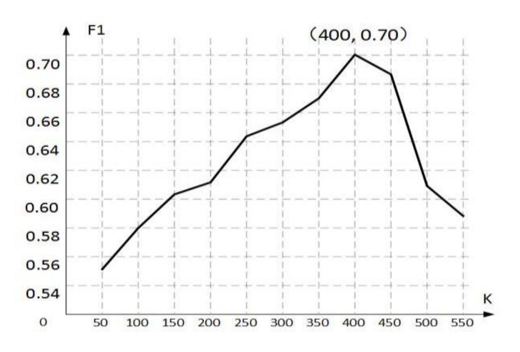


图 3　不同 K 值下准确率比较（五次计算取平均值）

从图 3 可以看出，经过多次计算取得平均值的情况下，采用相同的文本集和参数设定，主题数量设定为 400 时，F(1)为 0.70 最大。因此，在后续的 LDA 建模实验中，将主题数量设置为$k = 400$。

3.5　词语规模与运行时间实验对比

文本提出的 WMF_LDA 主题模型，在进行建模之前，首先根据词语语义将领域词与近义词进行统一化的映射表示。并根据中文文本的特点，将特定的名词和动词进行筛选构成新的文本集。从文本的语义结构和组织结构两方面进行语料库规模的压缩，可以降低词语数量，提高建模时间。下图表示了采用相同上述数据集，WMF_LDA 与传统 LDA 主题模型在词语数量与运行时间上的差异。

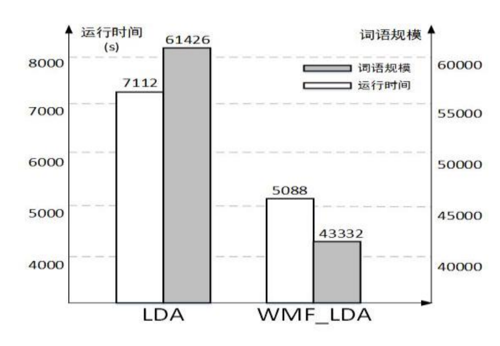


图 4　两种方法词语规模与运行时间比较（五次计算去平均值）

通过图 4 可以得知，本文提出的 WMF_LDA 在运行时间与词语规模上均优于传统 LDA 主题建模的方法。其中传统建模方法需要对 6 万多个词语建模，最终耗时 7000 多秒；本文提出的 WMF_LDA 对 4 万多个词语进行建模，最终耗时 4000 多秒。在词语规模与运行时间上均降低为原来的三分之二。

3.6　聚类准确率对比

本文采用经典的 K-means 聚类算法，以准确度 F1 值为评

价指标来衡量文本相似度计算效果。在效果对比时，本文采用传统 TF-IDF 方法、经典 LDA 方法与本文提出的 WMF_LDA 模型进行比较，并将本文提出的基于语义的词语合并与基于词性的词语筛选（WMF, Word Merging and Filtering）与传统 TF-IDF 相结合，一并作为对比实验。实验结果如表 3 所示。

表 3 不同方法下准确率比较(五次计算取平均值)

方法	准确度 F1 值
TF-IDF	60.1%
TF-IDF+WMF	61.8%
LDA	68.1%
WMF_LDA（本文方法）	72.5%

表 3 显示，本文提出的 WMF_LDA 方法在文本相似度计算与聚类准确度方面较传统 LDA 方法有明显的提升。同时，将本文提出的基于语义的词语合并与基于词性的词语筛选（word merging and filtering,WMF）应用于传统 TF-IDF 方法上，也可以获得一定程度的提升。这是因为本文提出的方法在文本语义结合和组成结构上，将对文本影响较小的词语筛选过滤，同时将能够体现文本领域特征的词语进行统一化映射，增强了文本的领域差异。

4 结束语

本文在分析传统 TF-IDF 和 LDA 计算文本相似度的基础上，提出 WMF_LDA 主题模型。其针对不同类型文本具有其特殊的领域词集的特点，根据词语的语义将同领域内或相近语义的词语映射到同一个词语表示上，提高了领域词的出现频率，并增强其对所属领域的代表和表示能力，在主题建模时通过词频的增加提高了文档在某个主题下的分布概率。实验结果表明，WMF_LDA 可以降低词语规模、减少主题建模时间，并提高文本聚类的准确率。

下一步的工作是：在词语语义的基础上，考虑代词、形容词、副词的作用，进一步挖掘文本中其他组成成分之间的结构和语义关系，以进一步从句子的角度挖掘相似度计算方式，并映射到整篇文本的相似度。

参考文献：

[1] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报, 2011, 34 (5): 856-864. (Huang Chenghui, Yin Jian, Hou Fang. A text similarity measurement combining word semantic information with TF-IDF method [J]. Chinese Journal of Computers, 2011, 34 (5): 856-864.)

[2] 徐浩广，王宁，刘佳明，等. 基于自然语言检索的综合相似度计算算法 [J]. 计算机系统应用, 2017, 26 (6): 170-175. (Xu Guanghao, Wang Ning, Liu Jiaming, *et al.* Comprehensive computation algorithm of similarity for natural language retrieval [J]. Computer Systems & Applications, 2017, 26 (6): 170-175.)

[3] Erkan G, Radev D R. LexRank: graph-based lexical centrality as salience in text summarization [J]. Journal of Artificial Intelligence Research, 2004 22 (1): 457-479.

[4] Theobald M, Siddharth J, Paepcke A. SpotSigs: robust and efficient near duplicate detection in large web collections [C]// Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2008: 563-570.

[5] 郭庆琳，李艳梅，唐琦. 基于 VSM 的文本相似度计算的研究 [J]. 计算机应用研究, 2008, 25 (11): 3256-3258. (Guo Qinglin, Li Yanmei, Tang Qi. The similarity computing of documents based on VSM [J]. Application Research of Computers, 2008, 25 (11): 3256-3258.)

[6] 苏小虎. 基于改进 VSM 的句子相似度研究 [J]. 计算机技术与发展, 2009, 19 (8): 113-116. (Su Xiaohu. research of sentence similarity based on improved VSM [J]. Computer Technology & Development, 2009, 19 (8): 113-116)

[7] 谷重阳，徐浩煜，周晗，等. 基于词汇语义信息的文本相似度计算 [J]. 计算机应用研究, 2018 (2): 391-395. (Gu Chongyang, Xu Haoyu, Zhou Han, *et al.* Text similarity computing based on lexical semantic information [J]. Application Research of Computers, 2018 (2): 391-395.)

[8] Blanco E, Dan M. A Semantic logic-based approach to determine textual similarity [J]. IEEE//ACM Trans on Audio Speech & Language Processing, 2015, 23 (4): 683-693.

[9] Atoum I, Otoom A. Efficient hybrid semantic text similarity using wordnet and a corpus [J]. International Journal of Advanced Computer Science & Applications, 2016, 7 (9) .

[10] 黄贤英，李沁东，刘英涛. 结合词性的短文本相似度算法及其在文本分类中的应用 [J]. 电讯技术, 2017, 57 (1): 78-82. (Huang Xianying, Li Qindong, Liu Yingtao. A grammatical category-combined short-text similarity algorithm and its application in text categorization [J]. Telecommunication Engineering, 2017, 57 (1): 78-82)

[11] 黄贤英，张金鹏，刘英涛，等. 基于词项语义映射的短文本相似度算法 [J]. 计算机工程与设计, 2015 (6): 1514-1518. (Huang Xianying, Zhang Jinpeng, Liu Yingtao, *et al.* Short text similarity algorithm based on term mapping with semantic [J]. Computer Engineering & Design, 2015 (6): 1514-1518)

[12] 黄江平，姬东鸿. 基于卷积网络的句子语义相似性模型 [J]. 华南理工大学学报：自然科学版, 2017, 45 (3): 68-75. (Huang Jiangping, Ji Donghong. Sentence semantic similarity model based on convolutional networks [J]. Journal of South China University of Technology: Social Science Edition, 2017, 45 (3): 68-75)

[13] Kenter T, Rijke M D. Short text similarity with word embeddings [C]// Proc of ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1411-1420.

[14] Kusner M J, Sun Y, Kolkin N I, *et al.* From word embeddings to document distances [C]// Proc of International Conference on International Conference on Machine Learning. 2015: 957-966.

[15] Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with siamese recurrent networks [C]// Proc of Repl4nlp Workshop at ACL. 2016.

[16] Kashyap A, Han L, Yus R, *et al.* Robust semantic text similarity using LSA, machine learning, and linguistic resources [J]. Language Resources & Evaluation, 2016, 50 (1): 125-161.

[17] 孙昌年, 郑诚, 夏青松. 基于 LDA 的中文文本相似度计算 [J]. 计算机技术与发展, 2013 (1): 217-220. (Sun Changnian, Zheng Cheng, Xia Qingsong. Chinese text similarity computing based on LDA [J]. Computer Technology & Development, 2013 (1): 217-220)

[18] 张超, 陈利, 李琼. 一种 PST_LDA 中文文本相似度计算方法 [J]. 计算机应用研究, 2016, 33 (2): 375-377. (Zhang Chao, Chen Li, Li Qiong, *et al.* Chinese text similarity algorithm based on PST_LDA [J]. Application Research of Computers, 2016, 33 (2): 375-377.)